Analysis and Prediction of Hydrogen Bonding in Protein–DNA Complexes Using Parallel Processors

GRAHAM CAMPBELL
Brookhaven National Laboratory, Upton, New York 11973

YUEFAN DENG, JAMES GLIMM, YUAN WANG, and QIQING YU
Department of Applied Mathematics and Statistics, SUNY at Stony Brook, Stony Brook, New York 11794-3600

MOISÉS EISENBERG* and ARTHUR GROLLMAN
Department of Pharmacological Sciences, SUNY at Stony Brook, Stony Brook, New York 11794-8651

Received 15 February 1995; accepted 17 August 1995

ABSTRACT

A number of essential biological functions are controlled by proteins that bind to specific sequences in genomic DNA. In this article we present a simplified model for analyzing DNA–protein interactions mediated exclusively by hydrogen bonds. Based on this model, an optimized algorithm for geometric pattern recognition was developed. The large number of local energy minima are efficiently screened by using a geometric approach to pattern matching based on a square-well potential. The second part of the algorithm represents a closed form solution for minimization based on a quadratic potential. A Monte Carlo method applied to a modified Lennard–Jones potential is used as a third step to rank DNA sequences in terms of pattern matching. Using protein structures derived from four DNA–protein complexes with three-dimensional coordinates established by X-ray diffraction analysis, all possible DNA sequences to which these proteins could bind were ranked in terms of binding energies. The algorithm predicts the correct DNA sequence when at least two hydrogen bonds per base pair are involved in binding to the protein, providing a partial solution to the three-dimensional docking problem. This study lays a framework for future refinements of the algorithm in which the number of assumptions made in the present analysis are reduced.

* Author to whom all correspondence should be addressed.
Editor's Note: Corrected proofs for this article were received on September 18, 1996.

CCC 0192-8651 / 96 / 151712-14

Introduction

Essential biological functions, including transcription, gene regulation, and the action of restriction endonucleases, are controlled by proteins that bind to specific nucleotide sequences in DNA.[1-3] Initial binding of a protein to DNA may be nonspecific; however, either molecule may undergo conformational change to form a stable sequence-specific complex in which the free energy of the complex is minimized.[4] Sequence recognition involves direct hydrogen bonding and van der Waals interactions between side chains of the protein and edges of base pairs exposed in the grooves of duplex DNA.[5] Electrostatic forces also are involved in stabilizing DNA–protein complexes, but make a lesser contribution to sequence specificity. Global structure of DNA, moreover, may depend on base sequence, and thereby influence contacts on the protein–DNA interface. Both protein and DNA may be subject to induced fit,[6] a conformational change on flexible domains which may occur upon binding to each other.

The secondary structure of proteins plays an important role in binding to DNA. Helices contain side chains that form hydrogen bonds with complementary DNA bases. Several motifs, including the α helix, helix-turn-helix, zinc finger, and leucine zipper, bind DNA in a sequence-specific manner.[7] Helices contain side chains that form hydrogen bonds with complementary DNA bases.

Proteins principally interact with DNA in the major groove which presents a greater diversity of hydrogen-bonding sites. Patterns of donors and acceptors are unique for each base pair.[5] Another general feature of DNA–protein complexes is that sequence-specific interactions occur at the amino end of the α helix which penetrates the major groove of DNA.[8] Protein side chains do not always recognize the same base pair although certain preferences are observed. As suggested by Seeman et al.,[5] the ability of these side chains to make bidentate interactions enhances their capability for sequence-specific recognition.

Energetic considerations relating to protein–DNA complexes have been reviewed by Berg and Von Hippel.[8] Rigid DNA-binding domains in the protein drive selective recognition by imposing H-bonding interactions that destabilize nonspecific protein–DNA complexes relative to complexes with the correct target sequence. Target sequences are distinguished from a vast number of competing nonspecific sites by means of significant free energy differences.[4,9]

The principles governing sequence-specific hydrogen bonding of proteins to DNA are supported by X-ray diffraction analysis of various DNA–protein complexes.[7,10] The relative contributions of H bonding, van der Waals interactions, and electrostatic forces to sequence-specific binding have not been clearly established. In this article we explore the hypothesis that H bonding is of overriding importance by formulating algorithms that would allow prediction of the DNA sequence(s) to which such a protein would bind. To test the extent to which global minimization of binding energy of hydrogen bonds is sufficient to predict specificity, we have analyzed four established protein–DNA complexes, posing the following question: given the H-bonding capability of a protein in a fixed conformation and allowing for the observed flexibility in DNA structure, can we predict those sequences in DNA most likely to be accommodated in a protein–DNA complex?

Since the combinatorial complexity of all hydrogen-bonding pairs is astronomical, our computations necessarily involve certain assumptions. The protein is held rigid and only amino acids engaged in hydrogen bonding to DNA are used for the analysis. Van der Waals forces and electrostatic interactions between positively charged groups on the protein and phosphates on DNA are excluded from the computations since these forces contribute primarily to free energy of stabilization as opposed to sequence specificity. Although water is present in some crystalline complexes and may contribute to stability and sequence specificity,[11] solvent molecules have been ignored as a first approximation.

This study presents a geometric pattern matching algorithm, implemented for parallel computers, that allows us to rank DNA sequences in terms of binding energy for hydrogen bonds involved in complex formation. The algorithm correctly predicts DNA sequences where at least two hydrogen bonds per base pair are involved in binding the protein. Error frequency is higher when only one hydrogen bond is involved. This algorithm provides a partial solution to the docking problem and lays the framework for future refinements in which fewer assumptions are made.

We conclude from our analysis that the contribution of hydrogen bonding to the specificity of DNA–protein duplexes is of overriding impor-

tance in that sequence specificity can be predicted, based on the tertiary structure of a particular protein and duplex DNA, allowing for the observed conformational flexibility of DNA.

Methods

OVERVIEW

Given a protein with a set P of potential binding sites (hydrogen-bond donors or acceptors), we try to predict which of these sites in P, and which sites on DNA base pairs, form active hydrogen bonds. Figure 1 schematically shows the geometry of a typical hydrogen bond.

For a system with N active hydrogen bonds whose associated donors and acceptors span distances r_i, $i = 1, 2, \ldots, N$, the total system binding energy can be expressed as the sum of a modified pairwise Lennard–Jones (LJ) potential[12]

$$V = \sum_{i=1}^{N} V_i(r_i, \gamma_i, \theta_i, \phi_i) \qquad (1)$$

where

$$V_i(r, \gamma, \theta, \phi) = \begin{cases} c_0\left[(\sigma/r)^{\alpha} - (\alpha/\beta)(\sigma/r)^{\beta}\right] & \text{if } r < r_0 \\ c_0\left[(\sigma/r)^{\alpha} - (\alpha/\beta)(\sigma/r)^{\beta}\right]\{[\cos(\gamma) & \\ \quad +3\cos(\theta)\cos(\phi)]/4\} & \text{if } r > r_0 \end{cases} \qquad (2)$$

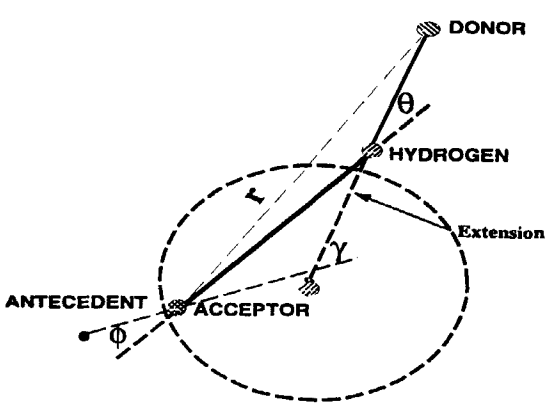

FIGURE 1. A typical hydrogen bond consists of a donor, acceptor, and a hydrogen atom. An acceptor antecedent with the acceptor defines the acceptor's dipole. Angles and distances are displayed.

and c_0 is a normalizing coefficient. Here $r = r_0$ is the solution to the equation

$$(\sigma/r)^{\alpha} - (\alpha/\beta)(\sigma/r)^{\beta} = 0.$$

σ is a weakly pair-dependent coefficient that determines the location of the minimum of the potential. The angles θ, ϕ, and γ are defined in Figure 1 and r is the distance between the donor and the acceptor. The form of the LJ potential $V_i(r, \gamma, \theta, \phi)$ used above for hydrogen bonds is an approximation. The radial LJ potential factor uses $\alpha = 6$ and $\beta = 4$, following Brunger,[12] reflecting the interaction between the two charged particles (donor and acceptor). The angular factor $\cos(\gamma) + 3\cos(\theta)\cos(\phi)$ is derived from the interaction energy between two point dipoles (the donor hydrogen and the antecedent acceptor).[13] Brunger[12] proposes a different form for the angle factor. Energy minimization is sensitive to the detailed form of the potential, including specifically the angular factors. It appears that our choice of angle factors may be an improvement over more commonly used choices. In the Discussion section, the quantitative accuracy of the potential [eq. (2)] is assessed relative to the data analyzed here.

In all of our studies, we normalize $c_0 = 1$ (without affecting the ranking results), and choose $\alpha = 6$, $\beta = 4$, and $\sigma = 2.87$ Å.

To illustrate the complexity of the problem, we estimate the number of possible hydrogen-bond pairs for the zinc finger–DNA complex.[14] The DNA structure is that of the canonical B-form.[15,16] There are 16 potential hydrogen-bond sites in the protein near the binding region, including those involved in complex formation. We vary the DNA base sequence and look for any possible match with at least eight hydrogen bonds. For a DNA sequence which is six base pairs in length and with four choices of base pair for each of these pairs, the number of possible DNA patterns is $4^6 = 4096$. Each base pair exposes in the major groove four positions where donors and acceptors can be present, two for each base. Only three of these locations have hydrogen-bonding atoms since guanine and adenine have two such atoms while cytosine and thymine have only one. Thus, each pattern of six base pairs has 4×6 sites. The number of ways to choose eight sites out of 24 is $\binom{24}{8}$. Similarly, for the protein, the number of ways to choose eight sites out of 16 is $\binom{16}{8}$. Thus, the total number of

possible matches is

$$4^6 \times \binom{24}{8} \times \binom{16}{8} \times 8! \approx 1.56 \times 10^{18}.$$

Assuming that we form these matches, since we know global minimization on each requires more than 10 s on a high-end computer workstation, the total time required would exceed 10^{10} years. This complexity must be reduced for a computationally feasible approach.

A hybrid method was developed to reduce the complexity. We treat the DNA and the protein as rigid bodies and determine a rigid body motion that will minimize the system's free energy. This hybrid method consists of three main elements: geometric hashing to obtain matching pairs, using the square-well potential; least-squares minimization of pairwise distances to rank the prediction given by the above step, which is then used to further filter out insignificant matches; and Monte Carlo searching to stochastically minimize the system's free energy defined by eq. (1) (Fig. 2). The first two elements determine rigid body motions that attempt to bring pairs of points into coincidence. To reconstruct the geometry of hydrogen bonds between pairs of points, we fix one end of the hydrogen bond in the data and compute an "ideal" position for the other by extending, from the first point, in the ideal direction predicted for a hydrogen bond. We then use this extended ideal point as the center of the square-well or quadratic potential when locating the other end of the hydrogen bond in the data.

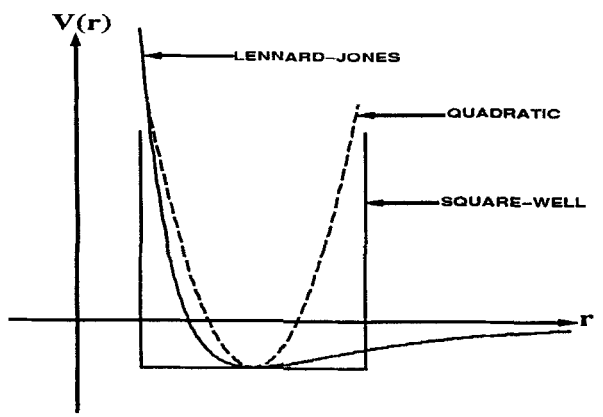

FIGURE 2. The stages of approximation for the Lennard–Jones potential. The sqaure-well potential is used to select interacting pairs and the quadratic potential is used to prioritize the pairs by minimizing the total geometric distance between the pairs that satisfy the square-well potential.

GEOMETRIC HASHING

The geometric hashing step is based on the fact that the LJ potential may be approximated by a square-well potential (Fig. 2). This method allows rapid screening to select only those pairs with the potential to be candidates for minimizing free energy. This idea was implemented previously by Nussinov and Wolfson.[17]

The geometric interpretation of the square-well potential is that we need only be concerned with sets of sites that can simultaneously be brought into coincidence within a distance, ε, by a rigid motion. In the following, we assume that equality of separation is always within tolerance of ε. These sets can be determined by examining the separation between points in one body and matching them with points that have a similar separation in the other body. First we create a list of distances separating pairs of points in the protein which can be used as a "lookup table" to determine rapidly all points separated by a given distance. Starting from two sites in the DNA placed as far apart as possible, we determine all pairs of points in the protein with an equal separation (within 2ε). For each of these, we determine all pairs of third points (one in the DNA, one in the protein) whose separations from each of the first two points agree. Thus, we have sets of triangles in the DNA that match (to within ε) a triangle in the protein. We now look for a matching tetrahedron using each triangle in the same manner. We have to insure, however, that the fourth point on each set lies on the same side of the plane of the first three (or else bringing them into coincidence would require a reflection as well as a rigid body motion) and this check must take into account the tolerance parameter in the case where all four points lie close to a single plane. The final match list for each triangle consists of all points that can form such a tetrahedron with the initial triangle.

This process creates a list of pairs of points which can be brought into coincidence, within the 2ε tolerance, by a rigid body motion of the DNA. This method can dramatically reduce complexity of the search, typically from the order of 10^{18} cases to 10^5. The pairs matched by this method usually include repetitions; these can be easily removed.

QUADRATIC APPROXIMATION

The next step is to compute the rigid body motion that minimizes the root mean squares (rms)

separation of the pairs of points created in the previous step. This transformation is defined by the rotation R and the translation b that minimizes the sum of the squared distances

$$S(R, b) = \sum_{i=1}^{N} \|Y_i - RX_i - b\|^2$$

between N pairs. A closed form solution for R can be derived as follows: define the 3×3 matrix

$$M = \sum_{i=1}^{N} (Y_i - \langle Y \rangle)^t (X_i - \langle X \rangle),$$

where $\langle Y \rangle$ is the average of the vector Y_is, etc. We factor M into the form $M = QDP$, where D is a diagonal matrix with diagonal elements $d_1 \geq d_2 \geq |d_3|$, Q and P are orthogonal matrices, and the determinant $|QP| = 1$. Then $R = P^tQ^t$ and $b = \langle Y \rangle - R\langle X \rangle$ is the closed form solution that minimizes $S(R, b)$.[18]

In considering a similar pattern recognition problem, Schwartz and Sharir[19] and Nussinov and Wolfson[17] presented a different solution for minimizing $S(R, b)$ by rigid motion. Their approach minimizes $S(R, b)$ by reflection and rigid motion; the solution given here allows only a simple rigid motion. The main difference between these solutions is the restriction $|QP| = 1$ imposed here.

Some of the matches constructed above give a conspicuously large rms pairwise separation; these matches are certain not to have the lowest energy. We use this information as a filter for reducing complexity.

We first reduce the maximal length match generated by geometric hashing to a specified length. We carry out this reduction by applying the rigid body motion computed above, then discarding the matched pair with the greatest separation. This is then repeated until a match of the desired length is achieved. The rms of the pairwise separations is computed; if it exceeds a specified value, this match is discarded. This provides another filter to reduce the complexity. This step now provides both a set of good candidates and represents a good starting point for the Monte Carlo search.

MONTE CARLO SCHEME

We have previously minimized the total system energy with the assumption that the pairwise interaction is quadratic (the standard distance geometry approach). We use this configuration as the starting point for minimization of the LJ potential which expresses a more realistic form of interaction in DNA–protein systems. A complex problem occurs when the LJ potential is used for pairwise interactions, namely, the existence of multiple local minima. Therefore, techniques such as Newton's method are not useful.

Other methods can be used to address this problem. We have tested two of these in our study: multiple starting-point Newton's method and very fast simulated reannealing (VFSR), developed by Ingber.[20] This algorithm was recently upgraded to include adaptivity; the revised package is called adaptive simulated annealing (ASA).†

In the first method, we decompose the 6-dimensional search domain (three for spatial variables, three for angular variables) centered at the quadratic minimum into subdomains of uniform sizes, then begin a search for a minimum by the conventional Newton's method starting from the center of each subdomain. The search in each subdomain is terminated when the subdomain boundary is reached. Thus, each subdomain will yield at most one minimum. At the end, we select a minimum among those obtained from the subdomains. The advantage of this method is its ease of implementation and its efficiency for problems with relatively low density of local minima.

The second method is adapted from VFSR.[20] Here we briefly explain the three major components of the algorithm: the probability function generation, the acceptance probability function generation, and the annealing schedule.

In the first component a new point, p^i_{k+1}, in an N-dimensional search domain (in our case, $N = 6$) is generated from an old point $p^i_k \in [X^i, Y^i]$ using a probability density function defined by the product of probability density functions for each parameter with density function, $g^i(z^i; T^i) = c[2(|z^i| + T^i)(1 + 1/T^i)]^{-1}$, in terms of a random variable $z^i \in [-1, 1]$, temperature parameters T^i, and renormalizing factor c. The formula relating old and new points and the random variables is

$$p^i_{k+i} = p^i_k + z^i(X^i - Y^i).$$

In component 2, the new point p^i_{k+i} is accepted based on the acceptance probability density. The cost functions, $C(p_{k+1}) - C(p_k)$, are compared us-

† This code can be obtained via anonymous ftp. L. Ingber, Adaptive Simulated Annealing (ASA), Lester Ingber Research, McLean, VA, 1993. The site address is ftp.alumni.caltech.edu. Relevant files are /pub/ingber/{ASA-shar, ASA.Zip}.

ing a uniform random number generator, $U \in (0, 1]$, in a Boltzmann test, i.e., we accept the new point if $\exp[-(C(p_{k+1}) - C(p_k))/T_{\text{cost}}] > U$ where T_{cost} is the temperature used for this test. Otherwise, the new point is discarded and the old is retained. The cost function in our case is the system's total energy.

An annealing schedule is derived in component 3. The annealing schedule for each parameter temperature, T^i, from a starting temperature T_0^i, is

$$T^i(k) = T_0^i \exp(-c^i k^{(1/D)})$$

where k is the iteration sweep counter and c^i and D are user-supplied parameters to adjust the annealing rate. The starting temperature T_0^i is normally chosen very high. To escape efficiently from local minima, VFSR has built in a concept of "reannealing." The efficiency of this method was tested by Ingber.[20] ‡ In VFSR, we artificially raise the annealing temperature to T_0^i each time one local minimum is located. This procedure in theory follows the "importance sampling" and we expect that it speeds up locating the global minimum (escaping from one local minimum). But we did not have any proof in the second half of the above statement.

Our tests comparing the multiple starting point Newton's method with the VFSR indicate that the latter is in general superior for our problem when compared to the former. We compare the "global" minima obtained by these two methods with the total system time held constant. After changing the parameters by selecting another DNA sequence in the search problem, we repeat the search and compare again. Although the multiple Newton's method is sometimes superior, on most occasions VFSR locates a lower energy minimum.

PARALLELIZATION

We used two distributed-memory MIMD parallel machines, the Intel iPSC/860 with 32 processors and the Intel Paragon with 56 processors. Three programming models for parallelizing this search were tested as follows:

1. all participating processors are used to search randomly in the entire 6-dimensional space (three angles and three coordinates);
2. the entire search space is decomposed into a number of subdomains and each processor is assigned to a subdomain for the search; and
3. one processor is used as the manager which hands out DNA patterns to the processors that are waiting to start a new search.

‡ Further tests on the efficiency of the simulated annealing technique on one processor and multiple processors are in progress. Interested readers may contact deng@ams.sunysb.edu for more information.

There are advantages and disadvantages to all three methods: we have found method 3 to be optimal. In method 1, when the number of processors is large, the probability of repeatedly searching the same location is not negligible and the parallel efficiency is lowered. In method 2, if the number of processors is large, the subdomain sizes become too small and the time needed to search will be small compared with the time needed to coordinate the processors. In method 3, a master-slave model, processor 0, is assigned as the master and the remaining processors as the slaves. The master keeps a dynamic list of the patterns that need to be processed. A slave requests a pattern by sending a message to the master. As soon as the pattern information has arrived, this slave starts the three-stage minimization (geometric hashing, least squares, and Monte Carlo). The slaves finishing one pattern will request another from the master and the process is repeated until the list in the master is empty. The master also performs minimization in its spare time.

It is obvious that different patterns require different times for minimization. Method 3 guarantees a natural load balance by dynamically dispatching computational loads. The communication cost is also minimal. When the number of processors is in the range of thousands, communication to the master processor might become a burden. In this case, we could dedicate two or more processors as intermediate masters in a hierarchal fashion. This way, the loss of a few processing nodes keeps thousands of slave nodes working efficiently.

COMPUTER PROGRAM RUNNING TIMES

The parallel speedup was essentially perfect. Computer times for geometric hashing and for the quadratic potential are minimal, and the Monte Carlo energy minimization with the LJ potential dominates the computation (with more than 95% of the total CPU cost). At this stage, the number of remaining patterns ranged from 10 to nearly 2000 per data set, depending on the number of base

pairs and protein binding sites considered. The computation took from 1–10 node-min to examine a possible bond matching configuration. A running time of up to 200 node-h is needed for the largest single data set with 12 base pairs and 16 protein binding sites.

DATA SETS

We have applied our algorithm to DNA/protein complexes available through the Brookhaven National Laboratory (BNL) Protein Data Bank (PDB) at the time this project was initiated.[21,22] Five complexes had complete atomic coordinates as determined by X-ray crystallography:

1. (P1ZAA) zinc finger zif268/DNA complex;
2. (P1LMB) lambda repressor-operator/DNA complex;
3. (3CRO) phage 434 CRO protein/DNA complex;
4. (2OR1) phage 434 repressor/DNA complex; and
5. (1HDD) homeodomain/DNA complex.

Of these, 1HDD has only one base pair with a single hydrogen bond and is therefore not useful for our calculations. We confined our study to the four remaining data sets.

HYDROGEN-BOND IDENTIFICATION

Normally, BNL PDB files do not contain atomic coordinates for hydrogen, due to experimental uncertainties associated with the small hydrogen mass, as well as to the convenience of reconstructing these coordinates from chemistry constraints. We used X-plor[12] to reinsert the hydrogen atoms. In the following analysis, hydrogen bonds involving water bridges are not considered.

We identify hydrogen bonds between the two components of the complex by geometric considerations. There are three main steps in identifying this process:

1. potential donor identification,
2. potential acceptor identification, and
3. potential acceptor antecedent identification.

These steps are executed sequentially for each hydrogen atom. In the following, D is donor, H is hydrogen, A is acceptor, and AA is acceptor antecedent.

In step 1, we identify as a potential donor any nitrogen, oxygen, or sulfur atom belonging to the same residue as the hydrogen that has a distance of 1.0 ± 0.15 Å from the hydrogen. For each potential donor identified in step 1, we execute step 2: scan the PDB data and look for nitrogen or oxygen atoms belonging to the other molecule that are not themselves donors and which have the following features:

1. The distance from the H is 2 ± 1.15 Å.
2. The angle D–H–A ($180° - \theta$ in Fig. 1) is greater than 90°.

For each donor–acceptor pair identified in step 2, we proceed to step 3 in which we identify the antecedent as the closest non-H atom in the same residue with the angle H–A–AA ($180 - \phi$ in Fig. 1) greater than 90°. If the two potential antecedents have nearly the same distance to the acceptor, a phantom antecedent is created at their midpoint.

A hydrogen bond is identified only when a D, H, A, and AA meeting the above criteria are identified and when the potential is negative. Although in principle there is a possibility of multiple acceptors for a single donor,[23] all bonds in the current data sets are uniquely defined by the above criteria.

SQUARE-WELL PARAMETER ESTIMATION

The square-well and the quadratic potential selection procedure make use of an ideal position for the matching site. We determine this position by extending the donor sites (or acceptor sites) on the DNA to establish an ideal acceptor site (or donor site). The extension is in the direction of D → H (or AA → A).

There are two parameters in the procedure: extension and tolerance. Extension is the length of an ideal hydrogen bond. Together with the above-defined ideal direction, we thereby define an ideal acceptor (or donor) location. Tolerance is the radius of the square-well potential. It is thus the distance from the ideal position for the matching site to the true site to be considered. The mean AH distance is 2.24 Å and the range is 1.6–2.99 Å. The least squares estimates of these parameters from our data analysis are 1.79 and 1.4 Å, respectively. We used extension parameters in the range 1.87–2.0 Å and tolerance parameters in the range 1.8–2.5 Å. Since the quadratic potential selection procedure is only used as a filter, these parameters do not

influence the final LJ potential and ranking as long as the tolerance is sufficiently large to accommodate the biologically correct sequence.

Results

HYDROGEN-BOND CHARACTERISTICS

For each amino acid, the potential hydrogen-bonding sites are known. Characteristics of the five data sets, as determined by this procedure, are given in Table I. This table includes only hydrogen bonds between DNA base pairs and the protein. The columns labeled H-bonded base pairs contain the number of DNA base pairs with hydrogen bonds to the protein. However, these base pairs typically are not contiguous and the column labeled DNA span represents the minimum length of DNA, in number of contiguous base pairs, that includes all the protein binding base pairs. The strong H bond column lists the number of hydrogen bonds whose negative LJ energy is at least 10% of the strongest LJ energy allowed. In the docking algorithm and in the following discussion, only strong hydrogen bonds are considered. The protein sites column is the number of potential binding sites on the protein included in the calculation.

In certain complexes, a considerable number of water molecules are buried deeply between protein and DNA. Water mediates binding through formation of water bridges and clusters involving several water molecules. In Table I, this is reflected in the number of base pairs lacking hydrogen bonds (column 4 minus column 2) and the number of base pairs with only a single hydrogen bond (column 3). As water-mediated binding is not included in our analysis, we only predict specificity for base pairs identified in column 2.

DNA CONFORMATION

Standard B-form DNA is flexible under physiological conditions and generally will experience many motions, including bending or twisting from its average B-form conformation. These potential deformations were not ignored in our calculations. Since our goal is a predictive model of specificity based on binding sites and geometry of the protein, we model the binding-induced conformational changes in the DNA. At the present stage in the development of our approach, we use the experimentally determined DNA shape as our basic structure. To extend this shape information to all other DNA sequences, we require a simple model based on a small number of degrees of freedom. We consider DNA as a chain with rigid links (individual base pairs in the geometry given by the average B-form DNA) and flexible joints. Thus, the shape parameters are those of $n - 1$ rigid motions so as to best fit n adjacent B-form base pairs into the experimental data. These $n - 1$ rigid motions can then be applied to an arbitrary sequence of n base pairs, generating a proposed binding conformation for a DNA segment of any base sequence.

We examined the accuracy of this fitting procedure for the 53 base pairs included in the segment of the DNA involved (see below). For each base pair, we computed the rms difference between the positions of the original atoms and the positions of the same atoms in the model bent to match the experimentally determined basic structure. The mean value of these rms differences is 0.21 Å with a standard deviation of 0.079 Å. The mean value of the magnitude of the LJ energy differences $|\Delta E| = |E_{PDB} - E_{model}|$ between the experimental and the bending model hydrogen bonds is 0.014 and the standard deviation is 0.08. Here, the energy scale is normalized so that the strongest (minimum) LJ hydrogen-bond energy is -1. The mean value of

TABLE I.
Hydrogen-Bond Mediated Binding.

Data Set	H-Bonded Base Pairs	Single H-Bond Base Pairs	DNA Span	Strong H Bonds	Weak H Bonds	Protein Sites
1	8	3	10	14	0	16
2	8	5	15	11	0	12
3	3	1	14	7	1	8
4	6	5	14	7	0	10
5	1	1	1	1	0	NA

the relative energy difference is $|(E_{PDB} - E_{model})/E_{PDB}| = 1.65\%$.

COMPARISON OF BINDING PREDICTION TO EXPERIMENTAL DATA

To determine the role played by hydrogen bonds in the specific protein binding to DNA, we applied our methods to the four protein/DNA complexes. Total free energy was calculated for each possible DNA sequence by the procedures described under Methods. We ordered the sequences according to this energy and examined the rank of the sequence selected by nature. If hydrogen bonds play an important role in pattern-DNA binding, we would expect the DNA sequence obtained from the BNL PDB to rank highly in this list.

We included in this calculation not only the sites in the protein that are involved in hydrogen bonding, but also all other hydrogen donors and acceptors present on the protein surface delimited by the binding sites, plus all nearest neighbors of these. These sites were selected as follows:

1. Identify potential donors, acceptors, and antecedents.
2. Select atoms identified in step 1 which are within the sphere of radius r centered at the hydrogen associated with a donor in the DNA, or centered at an acceptor in the DNA. The parameter r is selected so that there is a moderate number of protein sites. In particular, we select $r = 3$ Å for the first data set and $r = 3.4$ Å for the other three data sets. Even with this choice for data set 1, 16 sites are found in the protein compared to < 13 sites for the other three data sets. When the number of protein sites becomes large, data analysis is not computationally feasible since the number of configurations increases approximately by a factor of 4^{n_p}, where n_p is the number of additional protein sites considered.

To apply our algorithm, we must determine the number of matches to consider. Seeman et al.[5] argued that a single hydrogen bond is inadequate for uniquely identifying any particular base pair as this would lead to numerous degeneracies. For example, in data set 4, there are seven strong hydrogen bonds located on six base pairs. This data set has only one base pair with more than one hydrogen bond. Thus. the quality of the prediction results display sensitivity to the choice of hydrogen bonds and base pairs to be matched, and especially to the number of base pairs with only one hydrogen bond. We now explain how these choices are made in our analysis.

We have applied our code to all four data sets to examine as many base pairs as possible with at least two strong hydrogen bonds, allowing possible exceptions for a specific number of base pairs which have only one strong hydrogen bond.

When restricted to base pairs with at least two hydrogen bonds, our method identifies the biologically correct sequence as first ranked in all cases considered. Table II displays specificity prediction ranks, comparing experimental data to energy minimization ranking. In Table II, n_{BP} is the number of base pairs to be considered. (Among them, there are $n_{BP,1}$ base pairs for which there exists only one strong hydrogen bond and there are at least two strong hydrogen bonds on the remaining $n_{BP} - n_{BP,1}$ base pairs.) Considering only n_{BP} base pairs, we looked at matches of length n_{HB}, the total number of strong hydrogen bonds on these base pairs. Results of the calculation were then ordered by total LJ potential and the actual DNA sequence identified. Note that for a given sequence, the program will evaluate several different bond matching configurations. We ordered all bond matching configurations and determined the rank of the actual configuration. The sequence ranking is determined by first selecting the unique configuration which has the lowest potential

TABLE II.
Specificity Resulting from Multiple Hydrogen Bonds.

	$n_{BP,1} = 0$			$n_{BP,1} = 1$	
Data Set	Rank (Seq.)	n_{BP}	Rank (Seq.)	n_{BP}	BP Location
1	1	4	1	5	0
			1	5	5
			2	5	6
2	1	3	18	4	2
			22	4	5
			1	4	6
			1	4	7
			1	4	10
3	1	2	5	3	1
4	1	1	3	2	1
			1	2	3
			7	2	10
			5	2	12
			1	2	14

among configurations in the same sequence, then ordering these configurations. The ranking of the biological correct sequence is denoted by rank in Table II.

Columns under $n_{BP,1} = 0$ display our primary results for base pairs with at least two hydrogen bonds per pair, showing sequence rank 1 in all cases. The column $n_{BP,1} = 1$ lists all cases that include a single base pair for which there is a single strong hydrogen bond; all other base pairs have at least two hydrogen bonds. The column BP location gives the base pair having the single hydrogen bond. We do not display results for $n_{BP,1} > 1$, i.e., cases having at least two base pairs with a single hydrogen bond. There are many such cases, and the prediction results are relatively poor. The comparison clearly displays sensitivity to the number of base pairs having at most one hydrogen bond.

In Figure 3 we give empirical energy-rank curves for the four data sets, which are the curves connecting the 15 points (rank, energy) from the top 15 configurations in our output for each data set corresponding to the column $n_{BP,1} = 0$ in Table II. The height and the vertical position of the error bars in Figure 3 are discussed in the following section.

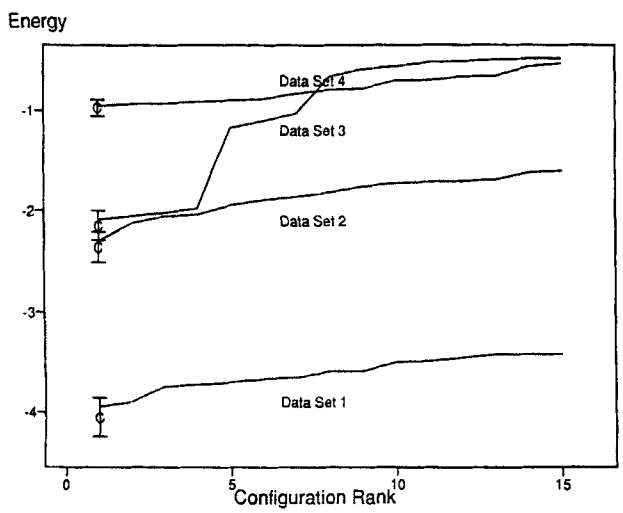

FIGURE 3. Energy versus ranked hydrogen bond configuration. Error bars are located at the biologically correct configuration. The size and location of the error bars represent the energy deviation due to bending ($N\langle \Delta E \rangle \pm 2$ SD) and C represents the predicted energy for the biologically correct unbent (B-form) sequence.

ERROR ESTIMATION IN PREDICTION

We consider the following types of error:

1. experimental error in BNL PBD data (approximately 0.1 Å);
2. bending model error (mean 0.2 Å);
3. potential energy formulation error (mean 0.38 Å);
4. computational error (due to convergence process in the Monte Carlo algorithm; estimated to be small enough not to affect our ranking conclusions).

We first discuss the error due to the bending model. Let $E_{lab,i}$, $i = 1, \ldots, k$, be the LJ energy of the strong hydrogen bonds from the PDB data, where k is the number of hydrogen bonds in a data set; and let $E_{model,i}$ be the LJ energy of the corresponding hydrogen bonds using our fitted DNA bending model. Let $\langle \Delta E \rangle = [\sum_{i=1}^{k}(E_{model,i} - E_{lab,i})]/k$, let rms denote $([\sum_{i=1}^{k}(E_{model,i} - E_{lab,i})^2]/k)^{1/2}$, and let SD denote $(\text{rms}^2 - \langle \Delta E \rangle^2)^{1/2}$.

We assume that deviations of energy per hydrogen bond due to the bending model are independently and identically distributed. Thus, an estimate of the mean deviation is the sample mean of the four data sets and an estimate of the standard deviation is the sample standard deviation for the four data sets. These are given in the last row of Table III. Since the LJ potential of a system under consideration is the sum of the LJ potentials of N individual hydrogen bonds, a natural estimate of the error in computing the LJ potential due to the bending approximation is $N\langle \Delta E \rangle \pm 2\sqrt{N}$ SD, where $\langle \Delta E \rangle$ and SD are from the last row of Table III. This analysis yields the error bars, as explained in the caption to Figure 3.

Table IV determines a ranking interval for the biologically correct sequence (see columns 3 and 4). The two end points of the prediction interval are determined by first subtracting $N\langle \Delta E \rangle \pm 2\sqrt{N}$ SD from the LJ energy predicted for the biologically correct sequence, using the bending model, then converting this energy interval into a ranking interval using the energy-rank relation displayed in Figure 3. The use of 2 SD in the energy interval gives these bounds an approximate 95% confidence interpretation for ranking errors due to the bending model. Other errors, such as corrections for full atomic interactions, are not represented in this estimate.

TABLE III.
Error Estimation (Energy / Bond) from Bending Model.

Data Set	$\langle \Delta E \rangle$	rms (Energy)	SD (Energy)	rms (Å)	SD (Å)
1	0.014815	0.024688	0.019749	0.1415	0.0335
2	0.016334	0.030606	0.025883	0.2020	0.0541
3	−0.007406	0.034150	0.033337	0.2697	0.0834
4	0.005735	0.043989	0.043614	0.2277	0.0678
Overall	0.010075	0.031541	0.029888	0.2147	0.0786

Provided our model assumptions are correct, one expects that the potential energy formulation error should be small, i.e., that coordinates obtained from the predicted sites for the biologically correct sequence should be essentially identical to the coordinates taken from experimental data. We compare the predicted coordinates to those of the laboratory data in Table V. Coordinates from the base pairs with at least two hydrogen bonds are used, i.e., data from the first three columns of Table II. Data set 4, with only one such base pair, is not rigidly bound by these two hydrogen bonds, generating a significant rms position error (2.3 Å) and is not used in computing the mean (average rms/base pair) error in Table V.

The deviation is reasonable when the number of base pairs is greater than 1. Table V is significant in that it tests not only the bending model error, but also the error in formulation of the LJ potential [eq. (2)] and, in fact, the error in restricting energy minimization to hydrogen bonds alone.

To reduce the Monte Carlo computational error while limiting computational costs, we use variable accuracy, or variable Monte Carlo running times, with the more accurate (longer) running times reserved for those cases with a relative chance of contributing to low ranking configurations. We then increased reanealing to 10 times the number used in the predictions for representative cases, including all data presented in Figure 3. In general, the result agreed within 0.03 in total energy and 0.2 Å in distance with lower bounds for the lower ranking cases important for our conclusions. We did observe a few instances of a large change in distance coordinates ($\geq$ 1 Å) with small change in energy, i.e., passage between distinct local minima; but these instances were above the energy and ranking threshold which would effect our conclusions.

PREDICTION ACCURACY PER BASE PAIR

For our top ranked configuration, and for at least two hydrogen bonds per base pair, no errors were found in our method for the data analyzed. To force errors and analyze their frequency, we consider the top two or top four ranking configurations and some base pairs with only one hydrogen bond. Within many such choices, we consider a representative sample for our error analysis. The most relevant measure of error is the error rate per base pair, as this parameter is dimensionless and does not grow artificially as the length of the comparison becomes longer. In Table VI, n_j is the number of base pairs for which there are exactly j hydrogen bonds in the data set. P_j is the proportion of correctly identified base pairs in the top four (or two) configurations of our results for base pairs with exactly j hydrogen bonds. In the P_j columns we give the sample mean with 2 SD (estimated, an approximate 95% confidence interval for P_j).

From Table VI, we note that the chance of correctly predicting the base pair which has at least two hydrogen bonds is approximately 90%,

TABLE IV.
Estimated Ranking Error due to Bending Model.

Data Set	N	Rank (Config.)	Rank (Seq.)
1	11	[1, 3]	[1, 2]
2	6	[1, 2]	[1, 2]
3	6	[1, 4]	[1, 1]
4	2	[1, 6]	[1, 2]

TABLE V.
Deviation of Predicted Coordinates from Laboratory Coordinates.

Data Set	1	2	3	Mean
rms (Å)	0.30	0.35	0.60	0.38
n_{BP}	4	3	2	1

TABLE VI.
Distribution of Correctly Identified Base Pairs.

Data Set	Top 2 Configurations (%)			TOP 4 Configurations (%)					
	P_3	P_2	P_1	P_3	P_2	P_1	n_3	n_2	n_1
1	100 ± 18	91 ± 21	50 ± 29	84 ± 13	95 ± 15	54 ± 20	1	4	3
2		93 ± 15	65 ± 22		87 ± 10	53 ± 16	0	3	5
3	75 ± 29		0 ± 50	83 ± 20		0 ± 20	2	0	1
4		88 ± 25	50 ± 19		84 ± 18	39 ± 13	0	1	5

and the chance of correctly identifying a base pair with exactly one hydrogen bond is about 50%. These numbers can be compared with the 25% accuracy obtained by random prediction (see Fig. 4).

Discussion

Using protein structures derived from DNA–protein complexes in which coordinates had been established by X-ray diffraction techniques, we have analyzed all possible DNA sequences to which these proteins might bind, ranking them in terms of binding energies based solely on contributions from hydrogen-bonded interactions. The accuracy of this analysis was found to depend on the number of donor and acceptor groups available on each base pair involved in the complex. When this number is two or more, the accuracy of prediction per base pair is 90%; when the number is one, this value is reduced to 50%. Van der Waals forces, which contribute significantly to the overall free energy of DNA–protein complexes, and water molecules, which can form hydrogen-bonded bridges between phosphate and base, phosphate and sugar, as well as between proteins and DNA,[24] were excluded from the analysis. Results of our study support the view (reviewed by Berg and von Hippel[8], and Steitz[7]) that hydrogen bonding between side chains of proteins and sites exposed in the major groove of DNA are critical determinants of recognition for proteins that bind in a sequence-specific manner to DNA.

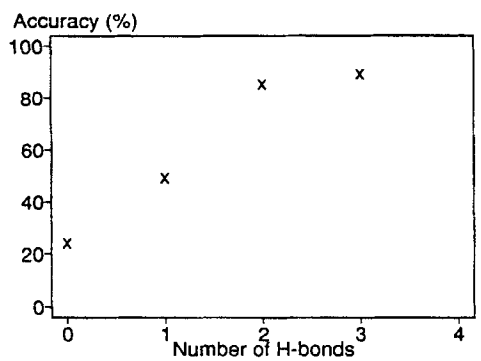

FIGURE 4. Approximately sigmoidal relation between accuracy of prediction and number of hydrogen bonds per base pair.

A number of simplifying assumptions were made in formulating algorithms used in this study. First, a potential energy function was constructed which involved only hydrogen bonds. Second, while bidentate bonds formed between side chains of proteins and DNA may be relevant to specificity,[5] we considered only one to one donor–acceptor type hydrogen bonding. Third, the protein was held completely rigid for the analysis while otherwise rigid base pairs in DNA were linked by flexible joints. Thus, the ability of B-form DNA to bend, kink, and otherwise adopt conformations containing variable tilt, roll, and twist parameters for individual base pairs, which can potentially alter the position and directionality of hydrogen bonds, was limited.[25] In all four complexes studied, the DNA helix was somewhat distorted relative to an average B-form when bound to the protein. We mimicked these same degrees of bending for our theoretically reconstructed sequences. Finally, the choice of an LJ potential for hydrogen bonds was necessarily approximate; moreover, the angular component used for this equation was based on a dipole–dipole interaction,[13] an assignment that proved superior to the angular factor used for molecular mechanics simulations.[12]

The complex pattern recognition problem involved in sequence-specific molecular recognition of duplex DNA by proteins was solved by significantly reducing the large number of combinatorial possibilities. This was accomplished by dividing the search algorithm into three components: geometric hashing, quadratic approximation, and the

Monte Carlo scheme. The latter was applied in conjunction with an LJ potential energy function. The first two steps in this process serve as "filters" to eliminate the vast majority of unwanted conformations. Schwartz and Sharrir[19] and Nussinov and Wolfson[17] applied these steps to an analogous problem, the latter group adding a reflection to rigid body motion in minimizing energies of protein–protein binding. These investigators did not utilize a Monte Carlo step in their computations.

In all cases where there were at least two hydrogen bonds for each base pair, the initial steps, involving application of square-well and quadratic potentials, yielded 100–350 configurations. When all base pairs containing a least one hydrogen bond were considered, 10^4 configurations remained after applying the first two steps of the algorithm. For the third computational step, approximately 30 node-min on the Paragon parallel computer were required to analyze each configuration. Thus, in less than 200 node-h, we were able to conduct computations involving a protein with 16 sites that bind to 12 base pairs in its cognate DNA.

Of the seven examples where ranking of sequences does not coincide with the biologically correct sequence (Table II), the predicted sequence in five cases differs from experimental data mainly at base pairs with a single hydrogen bond. The estimated error of coordinates of atoms for our model (< 0.4 Å) is comparable to the limits of atomic coordinate resolution of X-ray diffraction analysis.

In summary, we developed an algorithm that predicts intermolecular binding specificity between proteins and DNA based exclusively on minimization of energies contributed by hydrogen-bond interactions. This procedure is essentially a 3-dimensional docking algorithm. We validated this approach by predicting sequence specificity of DNA binding by four proteins for which the 3-dimensional structure of the protein–DNA complex was previously established by X-ray crystallography. In all cases, the sequence that ranks first in these calculations was identical with the one determined experimentally.

Future refinements of our algorithm will include other interactions, including van der Waals forces, such as those contributed by the methyl group of thymidine, steric factors, and water-mediated hydrogen bonds. The availability of additional experimentally determined data sets should help to further validate the model and simulations.

Acknowledgments

We thank J. Maizel for his useful comments and calling to our attention the work of Nussinov and Wolfson[17] in which a similar algorithm was used for the solution of the docking problem. This work was partially supported by the Applied Mathematics Subprogram of the U.S. Department of Energy DE-FG02-90ER25084 and under Contract DE-AC02-76CH00016. Y. D. was partially supported by NSF Grant DMS-9201581. J. G. was supported by the Army Research Office, Grant DA-AL03-89K0017 and through the Mathematical Sciences Institute of Cornell University under subcontract to the University at Stony Brook, ARO Contract DAAL03-92-G-0185, and the National Science Foundation Grants DMS-9201581 and DMS-9312098. Q. Y. was partially supported by NSF DMS-9202070 and M. E. and A. G. by NIH ES04068. We thank C. C. Chou for his help in several stages of the code development and tests.

References

1. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, Garland Publishing, New York, 1994.
2. C. Branden and J. Tooze, *Introduction to Protein Structure*, Garland Publishing, New York, 1991.
3. B. F. Luisi, W. X. Xu, Z. Otwinpwski, L. P. Freedman, K. R. Yamamoto, and P. B. Sigler, *Nature*, **352**, 497 (1991).
4. P. H. Von Hippel and O. G. Berg, *J. Biol. Chem.*, **264**, 675 (1989).
5. M. C. Seeman, J. M. Rosenberg, and A. Rich, *Proc. Natl. Acad. Sci. USA*, **73**, 804 (1976).
6. R. S. Spolar and T. Record, *Science*, **263**, 771 (1994).
7. T. A. Steitz, *Q. Rev Biophys.*, **23**, 205 (1990).
8. O. G. Berg and P. H. Von Hippel, *TIBS*, **13**, 207 (1988).
9. D. E. Draper, *Proc. Natl. Acad. Sci. USA*, **90**, 7429 (1993).
10. S. C. Harrison and A. K. Aggarwal, *Annu. Rev. Biochem.*, **59**, 933 (1990).
11. Z. Otwinowski, R. W. Schevitz, R. G. Zhang, C. L. Lawson, A. Joachimiak, R. Q. Marmorstein, B. F. Luisi, and P. B. Sigler, *Nature*, **335**, 321 (1988).
12. A. T. Brunger, *X-Plor Version 3.0. A System for X-Ray Crystallography and NMR*, Yale University Press, New Haven, 1992.
13. J. D. Jackson, *Classical Electrodynamics*, Wiley, New York, 1975.
14. N. P. Pavletich and C. O. Pabo, *Science*, **252**, 809 (1991).
15. S. Arnott and D. W. Hukins, *Biochem. Biophys. Res. Commun.*, **47**, 1504 (1972).
16. S. Arnott and D. W. Hukins, *J. Mol. Biol.*, **81**, 93 (1973).

17. R. Nussinov, and H. J. Wolfson, *Proc. Natl. Acad. Sci. USA*, **88**, 10495 (1991).
18. Y. Deng, J. Glimm, Q. Yu, and M. Eisenberg, *Appl. Math. Lett.*, **6**, 89 (1993).
19. J. T. Schwartz and M. Sharir, *Int. J. Robotics Res.*, **6**, 29 (1987).
20. L. Ingber, *Math Comput. Modelling*, **12**, 967 (1989).
21. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, Jr., E. F. M. Brice, M. D. Rodgers, J. R. Kennard, O. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
22. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, In *Crystallographic Databases—Information, Content, Software Systems, Scientific Applications*, F. H. Allen, G. Bergerhoff, and R. Sievers, Eds., Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987, p. 107.
23. G. A. Jeffrey and W. Saenger, In *Hydrogen Bonding in Biological Structures*, Springer–Verlag, New York, 1991, p. 20.
24. D. Vasilescu, J. Jaz, L. Packer, and B. Pullman, *Water and Ions in Biomolecular Systems*, Birkhauser Verlag, Berlin, 1990.
25. R. E. Dickerson and H. R. Drew, *Proc. Natl. Acad. Sci. USA*, **78**, 7318 (1981).